JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Terms of Use

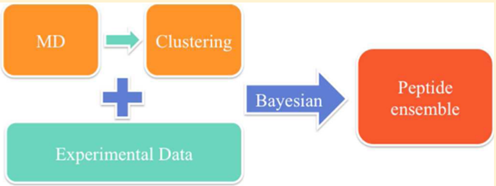
Peptide Conformation Analysis Using an Integrated Bayesian Approach

Xia Xiao,[†] Neville Kallenbach,[†] and Yingkai Zhang*,[†,‡]

[†]Department of Chemistry, New York University, New York, New York 10003, United States

[‡]NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai 200062, China

S Supporting Information



ABSTRACT: Unlike native proteins that are amenable to structural analysis at atomic resolution, unfolded proteins occupy a manifold of dynamically interconverting structures. Defining the conformations of unfolded proteins is of significant interest and importance, for folding studies and for understanding the properties of intrinsically disordered proteins. Short chain protein fragments, i.e., oligopeptides, provide an excellent test-bed in efforts to define the conformational ensemble of unfolded chains. Oligomers of alanine in particular have been extensively studied as minimalist models of the intrinsic conformational preferences of the peptide backbone. Even short alanine peptides occupy an ensemble of substates that are distinguished by small free energy differences, so that the problem of quantifying the conformational preferences of the backbone remains a fundamental challenge in protein biophysics. Here, we demonstrate an integrated computational-experimental-Bayesian approach to quantify the conformational ensembles of the model trialanine peptide in water. In this approach, peptide conformational substates are first determined objectively by clustering molecular dynamics snapshots based on both structural and dynamic information. Next, a set of spectroscopic data for each conformational substate is computed. Finally, a Bayesian statistical analysis of both experimentally measured spectroscopic data and computational results is carried out to provide a current best estimate of the substate population ensemble together with corresponding confidence intervals. This distribution of substates can be further systematically refined with additional high-quality experimental data and more accurate computational modeling. Using an experimental data set of NMR coupling constants, we have also applied this approach to characterize the conformation ensemble of trivaline in water.

1. INTRODUCTION

An emerging field in protein science is the study of intrinsically disordered proteins (IDPs),[1−3] which do not fold into well-defined 3D structures *in vitro* but are functional *in vivo*. IDPs appear to be abundant in nature—it has been predicted that about one-third of eukaryotic proteins contain extended disordered regions, including histone tails, α-synuclein, tau protein, p53, and BRCA1.[4] IDPs have been implicated in cellular functioning, especially in regulation and signaling.[5−7] Over 50 years ago, Tanford's sedimentation and viscosity measurements on denatured proteins led to a proposal of the random coil model for unfolded proteins,[8] which assumes that the polypeptide backbone freely samples all sterically allowed regions of the Ramachandran plot. In this view, unfolded proteins and peptides represent featureless "freely coiling" chains that occupy a multiplicity of conformations with very large associated backbone entropy. However, recently several lines of compelling spectroscopic evidence have converged to reveal that the backbone conformation of short unfolded peptides, including dipeptides and tripeptides, is structurally much more ordered than predicted by the random coil model.[9−14] The unfolded peptide backbone clearly has conformation preferences that are sequence and context dependent.[15−20] Thus, defining the conformations of peptides in unfolded states has become a problem of current interest and importance. Advances in this effort will enable construction of more accurate models of intrinsically disordered proteins, enable elucidation of fundamental principles of protein folding, and potentially help design novel functional peptide modulators of biological processes.

In contrast to folded globular proteins, which are routinely characterized at atomic resolution by X-ray crystallography and NMR spectroscopy, a comparably detailed characterization of unfolded peptides and proteins is much more challenging due to their multiplicity of conformational states and dynamic nature. For a given peptide, precise measurements can be made using a variety of spectroscopic methods, but data interpretation often requires *ad hoc* assumptions,[16,21−23] which introduce significant uncertainty and/or subjectivity into the final results. These make it difficult to utilize the complete set of available experimental data. For example, in one key study to determine polyproline II conformation propensities for a host−guest series of peptides $AcGGXGGNH_2$, only one experimental data set ($^3J_{\alpha N}$) was employed to fit each peptide to a two-state model, assuming that the experimental data are a weighted average of data from two representative basins, P_{II} and β.[24] On the other hand, limitations in sampling and force field accuracy

Received: May 19, 2014
Published: August 15, 2014

place direct determination of the relative stability of different conformational substates beyond the predictive power of molecular dynamics simulations using current force fields.[25−27]

One seminal attempt to overcome these difficulties is a combined molecular-dynamics/NMR (MD-NMR) approach developed by Graf et al.,[28] which aims to combine the accuracy of experimentally measured spectroscopic data with detailed information on the set of conformational substates provided by simulations. Taking the trialanine peptide in aqueous solution at 300 K as a model system, they experimentally measured a set of 15 J-coupling constants and carried out 100 ns of explicit water molecular dynamics simulations in parallel. Snapshots from the MD trajectory were assigned to three conformational substates (α, β, $\mathrm{P_{II}}$) based on the Ramachandran plot of the dihedral angles of the central residue. J-coupling values for each substate were calculated from corresponding Karplus equations. Finally, eight J-coupling constants for the central residue were employed to determine substate weights by performing a global fit to a three-state model, i.e., minimizing the difference between measured and calculated weight-average NMR parameters. This strategy represents a significant advance over studies that rely on either an experimental or computational approach alone. Nevertheless, this MD-NMR approach still has limitations: the analysis is restricted to a three-state model of an individual residue; the substates are predefined according to the Ramachandran plot, and only part of the experimental data set is used.

In this paper, we present an integrated computational-experimental-Bayesian framework (outlined in Figure 1) to characterize peptide conformational ensembles. This aims to overcome limitations in the published MD-NMR approach[28] by introducing two key features: (1) peptide conformational substates are assigned by clustering molecular dynamics snapshots based on both structural and dynamic information, rather than on subjectively defined rectangular regions of the Ramachandran plot; (2) a Bayesian statistical reweighting algorithm is used to provide an integrated analysis of both the experimental and computational data, which yields a current best estimate of substate populations with corresponding confidence intervals. This approach allows us to construct and assess multistate models of trialanine peptide in aqueous solution based on the full set of 15 measured J-couplings. Our results show that the two most dominant conformational substates of trialanine in water share the same polyproline II helix-like structure ($\mathrm{P_{II}}$) at its central residue, while differing at the C terminal residue. Our approach naturally allows for further systematic refinement using supplemental data sets and more accurate computational modeling of the relevant parameters.

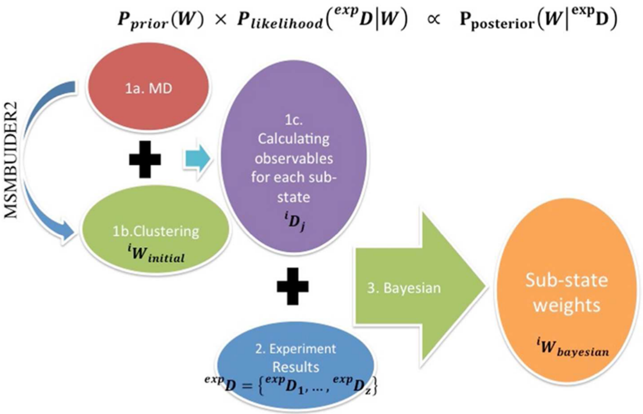

$P_{prior}(W) \times P_{likelihood}({}^{exp}D|W) \propto \mathrm{P_{posterior}}(W|{}^{\mathrm{exp}}\mathrm{D})$
MSMBUIDER2
1a. MD
1b.Clustering
${}^{i}W_{initial}$
1c. Calculating observables for each sub-state
${}^{i}D_{j}$
2. Experiment Results
${}^{exp}D = \{{}^{exp}D_1, \ldots, {}^{exp}D_z\}$
3. Bayesian
Sub-state weights
${}^{i}W_{bayesian}$

Figure 1. A schematic illustrating the Integrated Computational-Experimental-Bayesian approach.

2. METHODS

The central idea of the integrated computational-experimental-Bayesian framework to characterize peptide conformational propensities is illustrated in Figure 1. There are three steps in the computational stage: (1a) Extensive molecular dynamics simulations are carried out to generate an ensemble of peptide structure snapshots. (1b) MD snapshots are clustered into peptide conformational substates with a "divide-and-merge" approach based on both structural and dynamics information, allowing MD population weights of conformational substates to be calculated. (1c) For each conformational substate *i*, a set of values of spectroscopic data is computed. In the experimental stage, the key task is to obtain the corresponding experimentally measured spectroscopic data, either from the literature or by carrying out new experiments, or both. Finally, a Bayesian statistical algorithm is employed to provide an integrated analysis of both computational and experimental data, which yields a current best estimate of the substate populations as well as the corresponding confidence intervals.

In comparison with Graf's approach, two key components of this new integrated framework are the clustering step and the Bayesian statistical algorithm, which we discuss in more detail below.

2.1. Clustering.

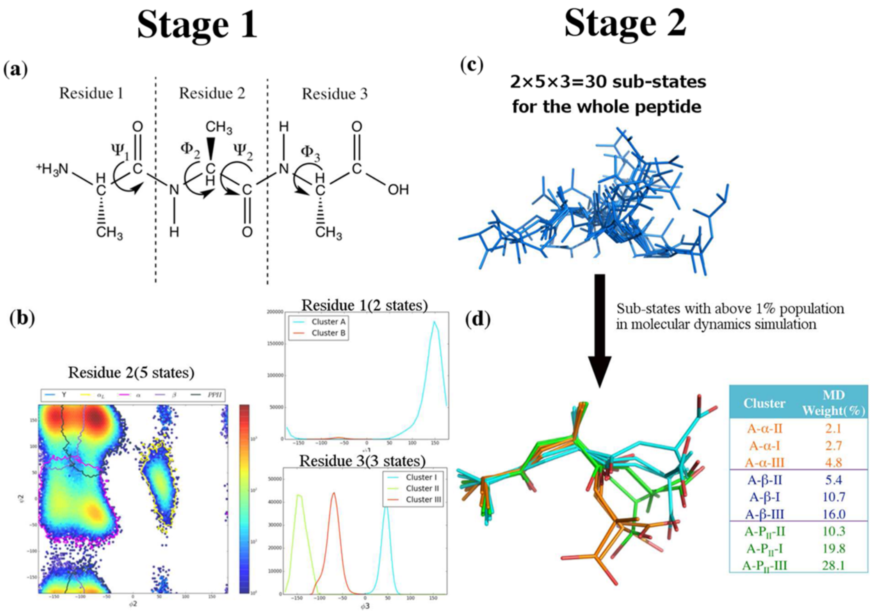
In most studies, peptide conformation substates are predefined with roughly rectangular regions of a Ramachandran plot. Typically, conformation assignment only considers backbone torsion angles of one single residue in a polypeptide.[16,26,28] Here, we present a more objective and robust method to define and assign peptide conformational substates, i.e., a "divide-and-merge" two-stage clustering approach. In the first stage, given a set of structural snapshots from molecular dynamics simulations, we use Markov state models to identify residue-based conformational macrostates based on both structural similarity and dynamics information by employing the program MSMBUILDER2.[29] Specifically, for each residue, MD trajectories are clustered into residue-based microstates using a hybrid k-centers k-medoids clustering algorithm[29] with the backbone RMSD as the structural similarity criteria. Then, kinetically related microstates are grouped together into residue-based macrostates using Perron Cluster Cluster Analysis (PCCA+).[30] In the second stage, these residue-based macrostates are merged to yield substates of the whole peptide.[31] For each substate *i*, its MD population weight ${}^{i}W_{\mathrm{md}}$ and a set of experimental observables ${}^{i}D$ can be computed.

2.2. Bayesian Statistical Weighting Algorithm.

With experimental data ${}^{\mathrm{exp}}D$ collected as well as the corresponding computed results ${}^{i}D$ for each conformational substate, a conventional approach to estimate substate population weights ${}^{i}W$, $i = 1,\ldots,n$, is to minimize an objective function, such as

$$\chi^2(W) = \sum_j [{}^{\mathrm{exp}}D_j - \sum_i {}^{i}W \times {}^{i}D_j]^2$$

in Graf's approach. This method tends to be limited to two to three substates of a single residue and fails to account for the uncertainty/error in either computed results or experimental data. In addition, slightly different objective functions can lead



Cluster	MD Weight(%)
A-α-II	2.1
A-α-I	2.7
A-α-III	4.8
A-β-II	5.4
A-β-I	10.7
A-β-III	16.0
A-P_{II}-II	10.3
A-P_{II}-I	19.8
A-P_{II}-III	28.1

Figure 2. "Divide-and-Merge" two-stage clustering of a trialanine MD trajectory simulated with Amber99SB forced field and TIP3P water. (a) Trialanine at pH = 2 with each residue labeled. (b) Stage 1, population distribution with residue-based clustering based on Markov state models. (c) Stage 2, residue-based macrostates are merged to yield a total of 30 substates for the whole peptide in principle, but only 22 substates existed in MD simulation. (d) Structures and populations of nine substates with above 1% population in MD simulation.

to distinct results, so that this minimization may not distinguish among several different solutions. In order to overcome the above limitations, here we employ a Bayesian statistical algorithm to provide an integrated analysis of both computational and experimental data[32,33] and determine conformational substate weights. In Bayesian inference, the belief in a hypothesis (H) is updated as additional evidence (E) is acquired by employing Bayes' rule:[34] $P(\mathrm{H|E}) = (P(\mathrm{E|H})\cdot P(\mathrm{H}))/P(\mathrm{E})$. The posterior probability of Bayesian inference $P(\mathrm{H|E})$, the updated belief in the hypothesis after incorporating additional evidence, is a function of two antecedents, a prior probability $P(\mathrm{H})$, which is the initial belief in the hypothesis, and a "likelihood function" $P(\mathrm{E|H})$, which is a conditional probability for evidence to be acquired given a hypothesis. $P(\mathrm{E})$ is the integrated likelihood of additional evidence, which is the same for all possible hypotheses being considered. In our characterization of peptide conformations, a set of substate weights **W** can be considered to be the hypothesis while experimental data ($^{\mathrm{exp}}\mathbf{D}$) are treated as additional evidence, which leads to the following formulation:

$$P_{\mathrm{posterior}}(\mathbf{W}|^{\mathrm{exp}}\mathbf{D}) = \frac{P_{\mathrm{prior}}(\mathbf{W})\cdot P_{\mathrm{likelihood}}(^{\mathrm{exp}}\mathbf{D}|\mathbf{W})}{\int P_{\mathrm{prior}}(\mathbf{W})\cdot P_{\mathrm{likelihood}}(^{\mathrm{exp}}\mathbf{D}|\mathbf{W})\ \mathrm{d}\mathbf{W}} \tag{1}$$

where $\mathbf{W} = \{^{1}W,...,^{n}W\}$ is the vector of weights for n substates subject to the constraint $\sum_{i=1}^{n} {}^{i}W = 1$ and ${}^{i}W \geq 0$; $^{\mathrm{exp}}\mathbf{D} = \{^{\mathrm{exp}}D_1,...,^{\mathrm{exp}}D_z\}$ is the vector of z experimental data.

Prior Distribution. $P_{\mathrm{prior}}(\mathbf{W})$ represents *a priori* knowledge about the weights of n conformational substates of the peptide. For each substate i, given its initial weight ${}^{i}W_{\mathrm{initial}}$ and its corresponding uncertainty $\sigma^2({}^{i}W_{\mathrm{initial}})$, a priori knowledge about the weight of this substate can be represented by a Gaussian distribution:

$$P_{\mathrm{prior}}({}^{i}W) = \frac{1}{\sqrt{2\pi\sigma^2({}^{i}W_{\mathrm{initial}})}}\ \mathrm{e}^{-({}^{i}W - {}^{i}W_{\mathrm{initial}})^2/2\sigma^2({}^{i}W_{\mathrm{initial}})} \tag{2}$$

Thus, the overall joint prior distribution can be expressed as

$$P_{\mathrm{prior}}(\mathbf{W}) = \prod_{i=1}^{n} P_{\mathrm{prior}}({}^{i}W) \tag{3}$$

with the constraints that $\sum_{i=1}^{ni} W = 1$ and ${}^{i}W \geq 0$. There are multiple ways to estimate values of ${}^{i}W_{\mathrm{initial}}$ and $\sigma({}^{i}W_{\mathrm{initial}})$ for eq 2. A straightforward approach is to employ information obtained from the MD simulations, the MD prior, which uses MD derived population weight ${}^{i}W_{\mathrm{md}}$ as the ${}^{i}W_{\mathrm{initial}}$. For $\sigma({}^{i}W_{\mathrm{initial}})$, we assign it an arbitrary large value of 20% when its uncertainty is not clear. As a control, if we do not use information from MD simulations, we calculate a simple random-coil (RC) based prior distribution, which assumes that each substate is equally populated, i.e., ${}^{i}W_{\mathrm{initial}} = 1/n$, where n is the total number of substates being considered.

Likelihood Function. $P_{\mathrm{likelihood}}(^{\mathrm{exp}}\mathbf{D}|\mathbf{W})$ represents the likelihood of observing the experimental data $^{\mathrm{exp}}\mathbf{D}$ given a certain substate weight **W**. For each given experimental observable $^{\mathrm{exp}}D_j$, the associated likelihood function can also be modeled with a Gaussian density function:

$$P_{\mathrm{likelihood}}(^{\mathrm{exp}}D_j|\mathbf{W}) = \frac{1}{\sqrt{2\pi(\sigma^2(^{\mathrm{exp}}D_j) + \sigma^2(^{\mathrm{comp}}D_j))}} \exp\left[-\frac{(^{\mathrm{exp}}D_j - {}^{\mathrm{comp}}D_j)^2}{2(\sigma^2(^{\mathrm{exp}}D_j) + \sigma^2(^{\mathrm{comp}}D_j))}\right] \tag{4}$$

$$^{\mathrm{comp}}D_j = \sum_{i=1}^{n} {}^{i}W \times {}^{i}D_j \tag{5}$$

Table 1. Nine-State Results for Trialanine Using the Amber99SB Force Field and TIP3P Water with Both MD Prior and Random Coil Prior[a]

		α			β			P_{II}			
		1	2	3	4	5	6	7	8	9	
Amber 99SB & TIP3P		A-α-II	A-α-I	A-α-III	A-β-II	A-β-I	A-β-III	A-P_{II}-II	A-P_{II}-I	A-P_{II}-III	χ^2
MD Prior	$W_{initial}(\sigma)$	2.1(20)	2.7(20)	4.8(20)	5.4(20)	10.7(20)	16(20)	10.3(20)	19.8(20)	28.1(20)	10.52
	$W_{bayesian}(\sigma)$	3.6(3.1)	2.4(2)	4(3.3)	1.5(1.4)	1.2(1.1)	1.8(1.7)	12.3(6.5)	4.3(3.2)	69.0(6.7)	3.17
RC Prior	$W_{initial}(\sigma)$	11.1(20)	11.1(20)	11.1(20)	11.1(20)	11.1(20)	11.1(20)	11.1(20)	11.1(20)	11.1(20)	14.63
	$W_{bayesian}(\sigma)$	3.8(3.2)	2.4(2.1)	4.6(3.6)	1.5(1.4)	1.2(1.1)	1.9(1.8)	13.7(7.1)	4.1(3.1)	66.9(7.1)	3.23

[a] $W_{initial}$ refers to *a priori* knowledge about weights of *n* conformation substates of the peptide. $W_{beysian}$ refers to the current best estimate of substate weights and their confidence interval. $\chi^2 = z^{-1}\sum_{j=1}^{z}({}^{exp}D_j - {}^{comp}D_j)^2/(\sigma^2({}^{exp}D_j)) + \sigma^2({}^{comp}D_j))$.

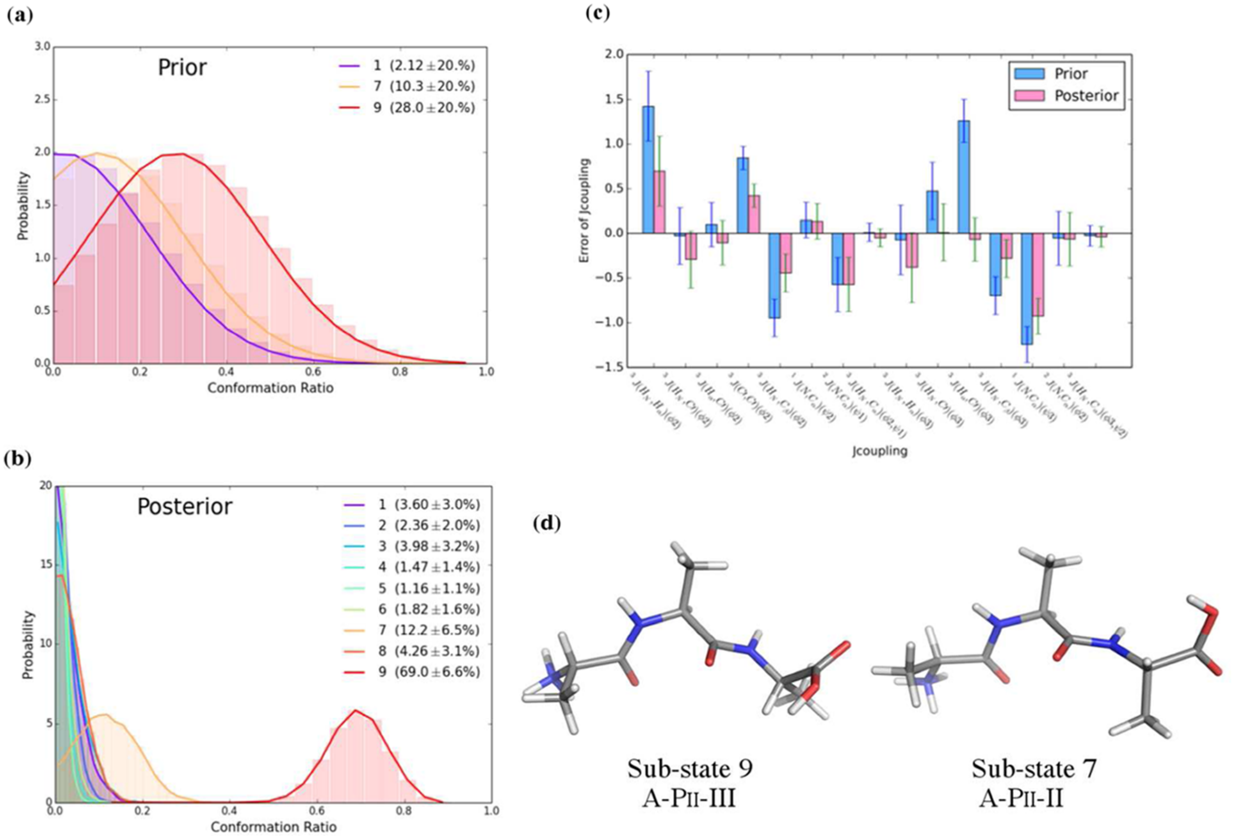


Figure 3. Nine-state results for trialanine with the Amber99SB force field and TIP3P water with the Bayesian algorithm and the MD prior: (a) Simulated prior distribution based on the MD prior. (b) Simulated posterior distribution of the final Bayesian model. (c) The differences between computed J-coupling constant and experimental J-coupling constant for both MD simulation and our approach. (d) Two dominant substates in the final Bayesian model.

where ${}^{i}D_j$ denotes the computed experimental observable *j* for the conformational substate *i*, $\sigma({}^{exp}D_j)$ refers to the uncertainty in the experimental measurement of each observable *j*, and $\sigma({}^{comp}D_j)$ is the error in theoretical prediction of the observable *j*. The overall joint likelihood function can be written as

$$P_{likelihood}({}^{\mathbf{exp}}\mathbf{D}|\mathbf{W}) = \prod_{j=1}^{z} P_{likelihood}({}^{exp}D_j|\mathbf{W}) \tag{6}$$

Once the prior distribution and the likelihood function are specified, the posterior distribution, $P_{posterior}(\mathbf{W}|{}^{\mathbf{exp}}\mathbf{D})$, our current best estimate of the conformational substate weights, is calculated using eq 1 by employing a Markov chain Monte Carlo (MCMC) algorithm.[35−37] The posterior distribution for each substate *i* can be computed by

$$P_{posterior}({}^{i}\mathbf{W}|{}^{exp}\mathbf{D}) = \int P_{posterior}(\mathbf{W}|{}^{exp}\mathbf{D})\ \mathrm{d}^{1}\mathbf{W}...\mathrm{d}^{i-1}\mathbf{W}\ \mathrm{d}^{i+1}\mathbf{W}...\mathrm{d}^{n}\mathbf{W} \tag{7}$$

The final Bayesian estimate of the weight and uncertainty of substate *i* can be computed by

$${}^{i}W_{\text{ bayesian}} = \int P_{posterior}({}^{i}W|{}^{\mathbf{exp}}\mathbf{D})\,{}^{i}W\ \mathrm{d}^{i}W \tag{8}$$

$$\sigma({}^{i}W_{\text{ bayesian}}) = \sqrt{\int P_{posterior}({}^{i}W|{}^{\mathbf{exp}}\mathbf{D})({}^{i}W - {}^{i}W_{\text{ bayesian}})^2\ \mathrm{d}^{i}W} \tag{9}$$

3. COMPUTATIONAL DETAILS

The initial simulation system was prepared by immersing peptide Ala_3 in a rectangular water box with a minimum solute−wall distance 15 Å, neutralized by adding one Cl^- counterion. Since the experiment was conducted at pH = 2, the N-terminus of the peptide would be protonated, as shown in Figure 2a. The AMBER 12[38] package with the Amber99SB force field[27,39−41] was used to perform classic MD simulations, and water molecules were described by the TIP3P[42] water model. Following multistep minimizations and MD equilibrations, a 200 ns NPT MD simulation was carried out. During the MD simulation, periodic boundary conditions were employed with a 10 Å cutoff for nonbonded interactions. Long-range electrostatic interactions were treated with the particle mesh Ewald (PME)[43,44] method. All bonds involving hydrogen atoms were constrained with the SHAKE[45] algorithm, and a time step of 2 fs was set. System temperature was controlled at 300 K with the Berendsen thermostat,[45] and the pressure was maintained at 1 atm. Snapshots were saved every 0.2 ps.

With 1 million snapshots from the MD simulations, a "divide-and-merge" two-stage clustering approach is employed to define and assign peptide conformational substates. In the first stage, for each residue, MD snapshots are clustered into residue-based microstates by employing the program MSMBUILDER2.[29] As illustrated in Figure 2b, the first residue of the trialanine peptide is clustered into two residue-based macrostates, the second into five macrostates, and the third into three macrostates. In the second stage, these residue-based macrostates are merged to yield a maximum of 2 × 5 × 3 = 30 substates in principle for the whole peptide (see Figure 2c), of which only 22 substates are populated sufficiently. We consider those substates that have >1% population in the MD simulations for further analysis, which includes nine conformational substates (see Figure 2d). The MD population for each substate *i* is then calculated based on the clustering results. The J-coupling constants for each snapshot are calculated from parametrized Karplus equations[46−49] (see Table S1), and average J-couplings constants ^iD are then computed for each substate i.

With the experimental data $^{\mathrm{exp}}\mathbf{D}$ collected for 15 J-coupling constants,[28] as listed in Table S2, and the corresponding computed values ^iD, we carry out a Bayesian statistical analysis, implemented with Python, to obtain the posterior distribution $P_{\mathrm{posterior}}(\mathbf{W}|^{\mathrm{exp}}\mathbf{D})$ in eq 1. The random walk Metropolis-Hastings algorithm[35−37] is used to sample the posterior distribution $P_{\mathrm{posterior}}(\mathbf{W}|^{\mathrm{exp}}\mathbf{D})$ in eq 1, and each posterior distribution has been sampled in 1 million steps. The random walk steps are obtained from a uniform distribution, and the step size of the random walk is adjusted to achieve a desired acceptance probability of 30%−70%.

4. RESULTS AND DISCUSSION

With the protocol described above, we have characterized a 9 substate ensemble for trialanine in aqueous solution with both MD prior (initial weights calculated from MD simulations) and RC prior (initial weights from a random-coil model, which assumes equal population among all substates). From Table 1, we can see that although the initial weights of the two prior distributions are very far apart, the Bayesian estimates of substate weights are consistent, and their final confidence intervals (σ) as well as the error in reproducing the experimental data (χ^2) are significantly smaller than employing

Table 2. 12-State Results for Trialanine with Amber99SB Force Field and TIP4PEW Water with MD Prior and Random Coil Prior

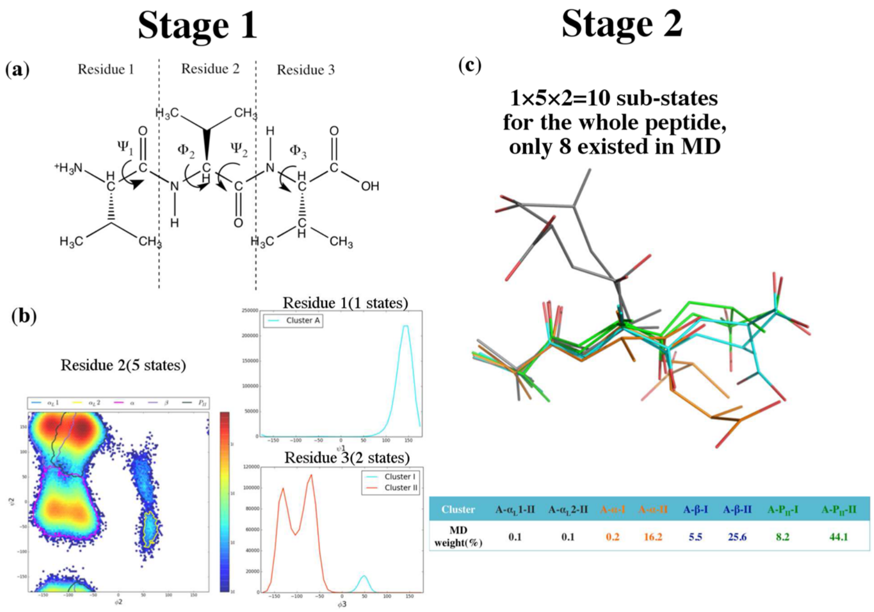
		α_L			α			β			P_{II}			
		10	11	12	1	2	3	4	5	6	7	8	9	
Amber 99SB & TIP4PEW		A-α_L-II	A-α_L-I	A-α_L-III	A-α-II	A-α-I	A-α-III	A-β-II	A-β-I	A-β-III	A-P_{II}-II	A-P_{II}-I	A-P_{II}-III	χ^2
MD Prior	$W_{\mathrm{initial}}(\sigma)$	2.2(20)	1.9(20)	2.9(20)	3.1(20)	2.1(20)	5.9(20)	5.4(20)	8.1(20)	14.9(20)	9.7(20)	17.3(20)	26.5(20)	9.78
	$W_{\mathrm{bayesian}}(\sigma)$	2.3(2)	1.6(1.4)	2.5(2.1)	3.4(2.9)	1.7(1.6)	3.8(3.2)	1.4(1.3)	1.1(1.1)	1.8(1.7)	10.6(6.3)	3.7(2.8)	66.0(6.6)	3.43
RC Prior	$W_{\mathrm{initial}}(\sigma)$	8.3(20)	8.3(20)	8.3(20)	8.3(20)	8.3(20)	8.3(20)	8.3(20)	8.3(20)	8.3(20)	8.3(20)	8.3(20)	8.3(20)	16.59
	$W_{\mathrm{bayesian}}(\sigma)$	2.2(1.9)	1.6(1.4)	2.8(2.3)	3.5(3)	1.7(1.6)	4.2(3.5)	1.4(1.4)	1.1(1.0)	1.7(1.6)	13.1(6.7)	3.6(2.8)	63.0(7.0)	3.49

Table 3. Five-State Results for Trialanine with Amber99SB Force Field and TIP3P Water with Both MD Prior and Random Coil Prior

Amber99SB & TIP3P		Y	α_L	α	β	P_{II}	χ^2
MD Prior	$W_{\text{initial}}(\sigma)$	0.024(20)	1.3(20)	9.8(20)	31.7(20)	57.2(20)	9.46
	$W_{\text{bayesian}}(\sigma)$	3.2(2.6)	3.5(2.7)	5.5(4.1)	2.0(1.8)	85.8(4.9)	2.10
RC Prior	$W_{\text{initial}}(\sigma)$	20.0(20)	20.0(20)	20.0(20)	20.0(20)	20.0(20)	20.78
	$W_{\text{bayesian}}(\sigma)$	4.0(3.0)	3.9(2.9)	7.6(5.1)	2.1(1.9)	82.4(5.6)	2.32

Table 4. Five-State Results for Trialanine with Amber99SB Force Field and TIP4PEW Water with MD Prior and Random Coil Prior

Amber99SB & TIP4PEW		Y	α_L	α	β	P_{II}	χ^2
MD Prior	$W_{\text{initial}}(\sigma)$	1.4(20)	6.9(20)	11.0(20)	28.0(20)	52.7(20)	8.99
	$W_{\text{bayesian}}(\sigma)$	3.3(2.6)	3.7(2.8)	6.5(4.6)	2.0(1.9)	84.5(5.2)	1.96
RC Prior	$W_{\text{initial}}(\sigma)$	20.0(20)	20.0(20)	20.0(20)	20.0(20)	20.0(20)	20.35
	$W_{\text{bayesian}}(\sigma)$	4.0(3.0)	4.0(2.9)	8.8(5.5)	2.1(1.9)	81.1(5.9)	2.17



Cluster	A-α_L1-II	A-α_L2-II	A-α-I	A-α-II	A-β-I	A-β-II	A-P_{II}-I	A-P_{II}-II
MD weight(%)	0.1	0.1	0.2	16.2	5.5	25.6	8.2	44.1

Figure 4. "Divide-and-merge" two-stage clustering of a trivaline MD trajectory simulated with the Amber99SB forced field and TIP3P water. (a) Trivaline at pH = 2 with each residue labeled. (b) Stage 1, population distribution with residue-based clustering based on Markov state models. (c) Stage 2, residue-based macrostates are merged to yield a total 10 substates for the whole peptide in principle, but only eight substates exist in the MD simulation.

Table 5. Eight-State Results for Trivaline with Amber99SB Force Field and TIP3P Water with MD Prior and Random Coil Prior

		α_L 1	α_L 2	α		β		P_{II}		
Amber 99SB & TIP3P		1	2	3	4	5	6	7	8	χ^2
MD Prior	$W_{\text{initial}}(\sigma)$	0.1(20)	0.1(20)	0.2(20)	16.2(20)	5.5(20)	25.6(20)	8.2(20)	44.1(20)	3.13
	$W_{\text{bayesian}}(\sigma)$	6.0(4.1)	6.1(4.1)	2.9(2.4)	19.2(7.6)	2.7(2.2)	8.8(5.5)	3.4(2.7)	51.1(7.1)	1.99
RC Prior	$W_{\text{initial}}(\sigma)$	12.5(20)	12.5(20)	12.5(20)	12.5(20)	12.5(20)	12.5(20)	12.5(20)	12.5(20)	7.13
	$W_{\text{bayesian}}(\sigma)$	6.4(4.3)	6.5(4.3)	2.9(2.5)	20.8(8)	2.8(2.4)	8.9(5.6)	4(3.1)	47.7(7.2)	2.05

Table 6. Five-State Results for Trivaline with Amber99SB Force Field and TIP3P Water with MD Prior and Random Coil Prior

Amber99SB & TIP3P		α_L 1	α_L2	α	β	P_{II}	χ^2
MD Prior	$W_{\text{initial}}(\sigma)$	0.1(20)	0.1(20)	16.4(20)	31.0(20)	52.3(20)	4.39
	$W_{\text{bayesian}}(\sigma)$	5.8(4.2)	7(4.5)	22.8(8.0)	10.5(5.9)	53.9(6.9)	2.46
RC Prior	$W_{\text{initial}}(\sigma)$	20.0(20)	20.0(20)	20.0(20)	20.0(20)	20.0(20)	9.46
	$W_{\text{bayesian}}(\sigma)$	6.2(4.4)	7.6(4.7)	25.7(8.2)	10.5(6)	50.0(6.9)	2.51

initial substate weights. Figure 3a and b illustrate a prior distribution from MD simulations as well as the posterior distribution of the final Bayesian model, and Figure 3c illustrates the significant reduction of error in reproducing the experimental data. These results clearly demonstrate the applicability and robustness of our integrated Bayesian approach.

In our nine substate ensemble, the A-P_{II}-III substate is the most dominant conformation, with a population of ~67%; the A-P_{II}-II substate is the second most dominant conformation, with a population of ~13%, as shown in Figure 3d. Both substates have the center amino acid in the P_{II} conformation but differ in the terminal dihedral angles. It should be noted that this level of characterization cannot be achieved by previous methods, which only focus on a single residue.

In order to assess the role of water models on our results, we have carried out molecular dynamics simulations using Amber99SB for the peptide and TIP4P-Ew[50] for water molecules, which previously have been shown to yield results in closer agreement with experimentally measured J-coupling data than the Amber99SB/TIP3P combination.[27] All other components in our computational and analysis protocol are the same as the above. From Table 2, we see that MD simulations with the Amber99SB/TIP4P-Ew force field yield 12 conformational substates with population levels above 1% after two-stage clustering. In comparison with results in Table 1, the three additional conformation substates have the central residue in the αL conformation. We have characterized the corresponding 12 substate ensemble for trialanine with both MD prior and RC prior, as shown in Table 2. Not only are the results for the different priors very consistent, the first and second major substates with populations of ~65% and ~12% are the same as in the nine-state model, which has populations of ~66% and ~13%, respectively. This further demonstrates the robustness of the integrated Bayesian approach.

To further examine its applicability and reliability, we have also carried out clustering and Bayesian analysis focusing on the central amino acid of Ala_3. The clustering results in five substates, as shown in Figure 2b. Only eight out of 15 experimental J-couplings (see Table S1 for those J-couplings labeled red) are related to dihedral angles of the center residue and were used to characterize this five-state ensemble. As shown in Tables 3 and 4, we can see that all results are very consistent, with the P_{II} conformation most dominant with a population of 86% ± 5%, 82% ± 6%, 84% ± 5%, and 81% ± 6%, respectively, for different priors and force fields. Meanwhile, all our results (Tables 1−4) consistently indicate that if focusing on the central residue, the α basin would be the second most populated (less than 10%) while the β conformation substate would be the least populated. It should be noted that Graf's three-state model[28] for the central residue of Ala_3 results in close to 0 population for the α conformation, which seems puzzling given the helix propensity of Ala.

Finally, we applied this integrated computational-experimental-Bayesian approach to characterize the conformational ensemble in trivaline in aqueous solution, as illustrated in Figure 4a. We carried out 200 ns molecular dynamics simulations using the Amber99SB/TIP3P force field, and snapshots were clustered into eight conformation substates as shown in Figure 4b and c. Using Graf's experimental data set of NMR coupling constants,[28] we determined an eight substate conformational ensemble for trivaline (Table 5) with a five-substate conformation ensemble for the central residue of trivaline (Table 6) using both MD and RC priors. The results are again very consistent despite employing different priors or different numbers of conformational substates. The most dominant conformation substate for trivaline has the center residue in the P_{II} conformation with a population ~49% ± 7%, much lower than that for trialanine.

5. SUMMARY

Conformational analysis of unfolded peptides is notoriously challenging, due to the intrinsically dynamic nature of the ensemble of accessible states that are distinguished by small free energy differences. Data from a variety of different spectroscopies including UVCD, VCD, Raman, and ROA have been used to demonstrate that there are in fact strong conformational preferences in unfolded states, modeled here by the trialanine and trivaline peptides in water. As pointed out in a detailed review by Adzhubei et al.,[51] the P_{II} conformation plays a major role in unfolded peptide structure. The main problem has been to quantify this or any other substate preference. In this work, we have demonstrated an integrated computational-experimental-Bayesian approach to characterize conformational ensembles. In comparison with previous methods, this integrated approach offers several novel attractive features: (i) It characterizes the whole chain rather than a single residue. (ii) It provides an objective and robust method to define and assign peptide conformational substates. (iii) It naturally includes uncertainty estimations, taking errors in both experimental data and computational results into account. (iv) Bayesian estimates of peptide conformational substates and their confidence intervals can be further systematically refined with additional high-quality experimental data and more accurate computational modeling, including more reliable force fields, more extensive sampling, and more accurate methods to compute experimental observables. Work along this line is currently in progress.

Here, we have applied this integrated approach to define the conformational ensembles of trialanine and trivaline in aqueous solution. Our results concur with other studies that disprove the random-coil model (a detailed review by Adzhubei et al.[51]) and indicate that P_{II} conformation is dominant in both tripeptides, to different degrees. One conclusion of the new approach is that the picture of a simple two-state distribution between β and P_{II} conformations is oversimplified. Our current analysis points to significantly lower populations of β structure than predicted by earlier studies.[18,28] The integrated strategy reported opens a way to quantitatively define the populations of conformation ensembles in unfolded peptides using a systematic and consistent procedure. Inclusion of different sequences and experimental data sets is currently being investigated.

■ ASSOCIATED CONTENT

Supporting Information

Tables S1, S2, and S3. This material is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: yingkai.zhang@nyu.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Y.Z. would like to acknowledge the support of NIH (R01-GM079223). We thank NYU-ITS and NYUAD for providing computational resources.

REFERENCES

(1) Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. R.; Hipps, K. W.; Ausio, J.; Nissen, M. S.; Reeves, R.; Kang, C. H.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, M.; Garner, E. C.; Obradovic, Z. *J. Mol. Graphics Modell.* **2001**, *19*, 26−59.

(2) Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. *Curr. Opin. Struct. Biol.* **2008**, *18*, 756−764.

(3) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197−208.

(4) Wright, P. E.; Dyson, H. J. *J. Mol. Biol.* **1999**, *293*, 321−331.

(5) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. *Biochemistry* **2002**, *41*, 6573−6582.

(6) Iakoucheva, L. M.; Brown, C. J.; Lawson, J. D.; Obradovic, Z.; Dunker, A. K. *J. Mol. Biol.* **2002**, *323*, 573−584.

(7) Iakoucheva, L. M.; Radivojac, P.; Brown, C. J.; O'Connor, T. R.; Sikes, J. G.; Obradovic, Z.; Dunker, A. K. *Nucleic Acids Res.* **2004**, *32*, 1037−1049.

(8) Tanford, C. *Adv. Protein Chem.* **1968**, *23*, 121−282.

(9) Poon, C. D.; Samulski, E. T.; Weise, C. F.; Weisshaar, J. C. *J. Am. Chem. Soc.* **2000**, *122*, 5642−5643.

(10) Woutersen, S.; Hamm, P. *J. Phys. Chem. B* **2000**, *104*, 11316−11320.

(11) Schweitzer-Stenner, R.; Eker, F.; Huang, Q.; Griebenow, K. *J. Am. Chem. Soc.* **2001**, *123*, 9628−9633.

(12) Woutersen, S.; Hamm, P. *J. Chem. Phys.* **2001**, *114*, 2727−2737.

(13) Grdadolnik, J.; Grdadolnik, S. G.; Avbelj, F. *J. Phys. Chem. B* **2008**, *112*, 2712−2718.

(14) Hagarman, A.; Measey, T. J.; Mathieu, D.; Schwalbe, H.; Schweitzer-Stenner, R. *J. Am. Chem. Soc.* **2010**, *132*, 540−551.

(15) Makowska, J.; Baginska, K.; Skwierawska, A.; Liwo, A.; Chmurzynski, L.; Scheraga, H. A. *Biopolymers* **2008**, *90*, 772−782.

(16) Toal, S.; Meral, D.; Verbaro, D.; Urbanc, B.; Schweitzer-Stenner, R. *J. Phys. Chem. B* **2013**, *117*, 3689−3706.

(17) He, L.; Navarro, A. E.; Shi, Z. S.; Kallenbach, N. R. *J. Am. Chem. Soc.* **2012**, *134*, 1571−1576.

(18) Grdadolnik, J.; Mohacek-Grosev, V.; Baldwin, R. L.; Avbelj, F. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1794−1798.

(19) Shi, Z. S.; Chen, K.; Liu, Z. G.; Kallenbach, N. R. *Chem. Rev.* **2006**, *106*, 1877−1897.

(20) Chen, K.; Liu, Z.; Zhou, C.; Shi, Z.; Kallenbach, N. R. *J. Am. Chem. Soc.* **2005**, *127*, 10146−10147.

(21) Shi, Z. S.; Olson, C. A.; Rose, G. D.; Baldwin, R. L.; Kallenbach, N. R. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 9190−9195.

(22) Schweitzer-Stenner, R.; Measey, T.; Kakalis, L.; Jordan, F.; Pizzanelli, S.; Forte, C.; Griebenow, K. *Biochemistry* **2007**, *46*, 1587−1596.

(23) Schweitzer-Stenner, R. *Mol. BioSyst.* **2012**, *8*, 122−133.

(24) Shi, Z.; Chen, K.; Liu, Z.; Ng, A.; Bracken, W. C.; Kallenbach, N. R. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 17964−17968.

(25) Gnanakaran, S.; Garcia, A. E. *J. Phys. Chem. B* **2003**, *107*, 12555−12557.

(26) Best, R. B.; Buchete, N. V.; Hummer, G. *Biophys. J.* **2008**, *95*, L7−L9.

(27) Nerenberg, P. S.; Head-Gordon, T. *J. Chem. Theory Comput.* **2011**, *7*, 1220−1230.

(28) Graf, J.; Nguyen, P. H.; Stock, G.; Schwalbe, H. *J. Am. Chem. Soc.* **2007**, *129*, 1179−1189.

(29) Beauchamp, K. A.; Bowman, G. R.; Lane, T. J.; Maibaum, L.; Haque, I. S.; Pande, V. S. *J. Chem. Theory Comput.* **2011**, *7*, 3412−3419.

(30) Deuflhard, P.; Weber, M. *Linear Algebra Appl.* **2005**, *398*, 161−184.

(31) Schultheis, V.; Hirschberger, T.; Carstens, H.; Tavan, P. *J. Chem. Theory Comput.* **2005**, *1*, 515−526.

(32) Fisher, C. K.; Huang, A.; Stultz, C. M. *J. Am. Chem. Soc.* **2010**, *132*, 14919−14927.

(33) Fisher, C. K.; Stultz, C. M. *Curr. Opin. Struct. Biol.* **2011**, *21*, 426−431.

(34) Bolstad, W. M. *Introduction to Bayesian Statistics*, 2nd ed.; John Wiley: Hoboken, NJ, 2007; p 464.

(35) Chib, S.; Greenberg, E. *Am. Stat.* **1995**, *49*, 327−335.

(36) Hastings, W. K. *Biometrika* **1970**, *57*, 97−109.

(37) Metropolis, N.; Ulam, S. *J. Am. Stat. Assoc.* **1949**, *44*, 335−341.

(38) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *Amber12*; *AmberTools 13*; University of California: San Francisco, CA, 2012.

(39) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712−725.

(40) Junmei, W.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.

(41) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(42) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(43) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(44) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(45) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.

(46) Karplus, M. *J. Chem. Phys.* **1959**, *30*, 11−15.

(47) Hu, J. S.; Bax, A. *J. Am. Chem. Soc.* **1997**, *119*, 6360−6368.

(48) Ding, K. Y.; Gronenborn, A. M. *J. Am. Chem. Soc.* **2004**, *126*, 6232−6233.

(49) Wirmer, J.; Schwalbe, H. *J. Biomol. NMR* **2002**, *23*, 47−55.

(50) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665−9678.

(51) Adzhubei, A. A.; Sternberg, M. J.; Makarov, A. A. *J. Mol. Biol.* **2013**, *425*, 2100−2132.